EXHIBIT B
(Part 1 of 3)

(54) Title: NOVEL OLIGONUCLEOTIDE ARRAYS AND THEIR USE FOR SORTING, ISOLATING, SEQUENCING, AND MANIPULATING NUCLEIC ACIDS

(57) Abstract

The present invention relates to new oligonucleotide arrays and methods of using oligonucleotide arrays. Binary oligonucleotide arrays, having binary oligonucleotides characterized by a constant nucleotide sequence adjacent to a variable nucleotide sequence, are used for sorting and surveying nucleic acid strands. Oligonucleotide arrays are used for sorting mixtures of nucleic acid strands, making immobilized partial copies of nucleic strands, ligating strands, or introducing site directed mutations into strands. Information is obtained for determining the sequence of a nucleic acid strand, alone or in a mixture, by generating partials of the strand and, for groups of partials having the same terminal variable oligonucleotide, separately determining the presence and sequence of all variable oligonucleotides. Arrays are also used to order previously sequenced nucleic acid fragments and to allocate ordered allelic fragments to chromosomal linkage groups.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	FR	France	MR	Mauritania
AU	Australia	GA	Gabon	MW	Malawi
BB	Barbados	GB	United Kingdom	NL	Netherlands
BE	Belgium	GN	Guinea	NO	Norway
BF	Burkina Faso	GR	Greece	NZ	New Zealand
BG	Bulgaria	HU	Hungary	PL	Poland
BJ	Benin	IE	Ireland	PT	Portugal
BR	Brazil	IT	Italy	RO	Romania
CA	Canada	JP	Japan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic	SD	Sudan
CG	Congo		of Korea	SE	Sweden
CH	Switzerland	KR	Republic of Korea	SK	Slovak Republic
CI	Côte d'Ivoire	KZ	Kazakhstan	SN	Senegal
CM	Cameroon	LI	Liechtenstein	SU	Soviet Union
CS	Czechoslovakia	LK	Sri Lanka	TD	Chad
CZ	Czech Republic	LU	Luxembourg	TG	Togo
DE	Germany	MC	Monaco	UA	Ukraine
DK	Denmark	MG	Madagascar	US	United States of America
ES	Spain	ML	Mali	VN	Viet Nam
FI	Finland	MN	Mongolia		

1

NOVEL OLIGONUCLEOTIDE ARRAYS AND THEIR USE FOR SORTING, ISOLATING, SEQUENCING, AND MANIPULATING NUCLEIC ACIDS

Field of the Invention

This invention is in the field of sorting, isolating, sequencing, and manipulating nucleic acids.

Background of the Invention

Ordered arrays of oligonucleotides ("oligos") immobilized on a solid support have been proposed for sequencing DNA fragments. It has been recognized that hybridization of a cloned single-stranded DNA fragment to all possible oligo probes of a given length can identify the corresponding, complementary oligo segments that are present somewhere in the fragment, and that this information can sometimes be used to determine the DNA sequence. Use of arrays can greatly facilitate the surveying of a DNA fragment's oligo segments.

In an oligonucleotide array each oligo probe is immobilized on a solid support at a different predetermined position. The array allows one to simultaneously survey all the oligo segments in a DNA fragment strand. Many copies of the strand are required, of course. Ideally, surveying is carried out under conditions to ensure that only perfectly matched hybrids will form. Oligo segments present in the strand can be identified by determining those positions in the array where hybridization occurs. The nucleotide sequence of the DNA sometimes can be ascertained by ordering the identified oligo segments in an overlapping fashion. For every identified oligo segment, there must be another oligo segment whose sequence overlaps it by all but one nucleotide. The entire sequence of the DNA strand can be represented by a series of overlapping oligos, each of equal length, and each located one nucleotide further along the sequence. As long as every overlap is unique, all of the iden-tified oligos can be assembled into a contiguous sequence block.

There is an important limitation to sequencing by known surveying techniques. As relatively longer DNA strands are surveyed, there is an increasing probability that more than two

-2-

identified oligos will share the same overlapping sequence, i.e.,
the overlap is not unique. When this occurs, the sequence of the
DNA cannot be unambiguously determined. Instead of one con-
tiguous sequence block that contains the entire DNA sequence, the
oligos can only be assembled into a number of smaller sequence
blocks, whose order is not known.

Summary of the Invention

We have invented new oligonucleotide arrays and methods of
using them.

A "binary array" according to the invention contains
immobilized oligos comprised of two sequence segments of prede-
termined length, one variable and the other constant. The
constant segment is the same in every oligo of the array. The
variable segments can vary both in sequence and length. Binary
arrays have advantages compared with ordinary arrays: (1) they
can be used to sort strands according to their terminal sequen-
ces, so that each strand binds to a fixed location (an address)
within the array; (2) longer oligos can be used on an array of a
given size, thereby increasing the selectivity of hybridization;
this allows strands to be sorted according to the identity of
internal oligo segments adjacent to a particular constant
sequence (such as a segment adjacent to a recognition site for a
particular restriction endonuclease), and this allows strands to
be surveyed for the presence of signature oligos that contain a
constant segment in addition to a variable segment; (3) universal
sequences, such as priming sites, can be introduced into the
termini of sorted strands using the binary arrays, thereby
enabling the strands' specific amplification without synthesizing
primers specific for each strand, and without knowledge of each
strand's terminal sequences; and (4) the specificity of hybrid-
ization during surveying can be increased by coupling hybridiza-
tion to a ligation event that discriminates against terminal
basepair mismatches.

A "sectioned array" as used herein is one divided into
sections, so that every individual area is mechanically separated

-3-

from all other areas, such as, for example, a depression on the
surface, or a "well". The areas have different oligos immobi-
lized thereon. A sectioned array allows many reactions to be
performed simultaneously, both on the surface of the solid
support and in solution, without mixing the products of different
reactions. The reactions occurring in different wells are highly
specific due to the nucleotide sequence of the immobilized oligo.
A large number of sortings and manipulations of nucleic acids can
be carried out in parallel, by amplifying or modifying only those
nucleic acids in each well that are perfectly hybridized to the
immobilized oligos. Nucleic acids prepared on a sectioned array
can be transferred to other arrays (replicated) by direct blot-
ting of the wells' contents (printing), without mixing the
contents of different wells of the same array. Furthermore, the
presence of individual sections in arrays allows multiple re-
hybridizations of bound nucleic acids to be performed, resulting
in a significant increase in hybridization specificity. It is
particularly advantageous according to this invention to use a
binary array that is sectioned.

Our invention includes methods of using sectioned arrays to
sort mixtures of nucleic acid strands, either RNA or DNA. As
used herein, "strand" means not just a single strand, but multi-
ple copies thereof; and "mixture of strands" means a mixture of
copies of different strands no matter how many copies of each are
present. Similarly "fragment" refers to multiple copies thereof,
and "mixture of fragments" means a mixture of copies of different
fragments. The methods include sorting strands either according
to their terminal oligo segments (3'-terminal or 5'-terminal), or
according to their internal oligo segments on a binary array.
Before or after sorting, universal priming region(s) can be added
to the strands' termini to enable amplification. Binary sec-
tioned arrays for sorting according to strands' terminal sequen-
ces ("terminal sequence sorting arrays") can be comprehensive. A
"comprehensive array" is one wherein any possible strand will
hybridize to at least one immobilized oligo. This type of
sorting is particularly useful for preparing comprehensive
libraries of fragments of a large genome. For example, in one

-4-

embodiment of the invention, strands of restriction fragments
have their restriction sites restored and are sorted on a binary
array. That array contains immobilized oligos whose constant
segments contain the sequence complementary to the restriction
site, and an adjacent variable segment. The array is complete,
containing all variable sequences of each type in separate areas.

Our invention also includes using sectioned arrays for
preparing every possible partial copy of a strand or a group of
strands. The term "partial" refers to multiple copies thereof.
Partials are prepared by either of the following methods: (1)
terminal sorting on a binary sectioned array of a mixture of all
possible partial strands generated by random degradation of a
parental strand; or (2) generation of partials directly on an
array, through the sorting on an ordinary sectioned array of
parental strands according to the identity of their internal
oligo sequences, followed by the synthesis of partial copies of
each parental strand by enzymatic extension of the immobilized
oligos utilizing the hybridized parental strands as templates.
In either case, generated partials correspond to a parental
strand whose 3' or 5' end is truncated to all possible extents
(at the "variable" end of the partial), and whose other end is
preserved (at the "fixed" end of the partial). These are "one-
sided partials." Unless otherwise indicated the word "partial"
is used herein to refer to one-sided partials.

Our invention also includes methods of using oligo arrays to
obtain oligo information as part of a process for determining the
nucleotide sequence of a long nucleic acid strand, or of many
nucleic acid strands in an unknown mixture. A complete set of
one-sided partials of the strand or strands is prepared on a
sectioned array, and the oligo content of the partial strands in
each well of the array is separately surveyed (i.e. each group of
partials sharing the same oligo at the partials' variable end is
surveyed).

Our invention also includes methods of using oligo arrays
for ordering previously sequenced fragments from a first restric-
tion digest of a large nucleic acid or even a genome.

–5–

Our invention also includes methods of using oligo arrays for allocating sequenced and ordered allelic fragments into their chromosomal linkage groups.

Our invention also includes a method of using binary arrays for surveying the oligos contained in strands or their partials. This method provides improved comprehensive surveys over the conventional surveying of oligos on an ordinary array.

Brief Description of the Drawings

Figure 1 shows a binary array.

Figure 1a shows an oligo immobilized in an area of a binary array.

Figure 2 shows a sectioned array having depressions.

Figure 2a shows a well of a sectioned array.

Figure 3 shows addition of a lattice to a support to make a sectioned array.

Figure 4 shows an example of sorting and amplification of restriction fragments on a sectioned binary array.

Figure 5 shows an example of preparing partials on a sectioned ordinary array.

Figure 6 shows, schematically, the order of steps for sequencing a complete genome.

Figure 7 shows, schematically, the use of a sheet with a number of miniature survey arrays for simultaneous surveying every well in a partialing array.

Figures 8 to 11 show examples of the determination of nucleotide sequences from indexed address sets obtained from analysis of mixtures of strands.

Detailed Description of the Invention

I.   Oligonucleotide arrays

As used herein an "oligonucleotide array" is an array of regularly situated areas on a solid support wherein different oligos are immobilized, typically by covalent linkage. Each area contains a different oligo whose location is predetermined.

-6-

Arrays can be classified by the composition of their immobilized oligos. "Ordinary arrays" contain oligos comprised entirely of "variable segments". Every position of the oligo sequence in such a segment can be occupied by any one of the four commonly occurring nucleotides.

Comprehensive ordinary arrays are those wherein any segment of any possible strand will hybridize perfectly to the length of one or more immobilized oligos so that no strand is lost.

Binary arrays differ from ordinary arrays. A binary array is illustrated in Figures 1 and 1a. Figure 1 shows a substrate or support 1 having immobilized thereon an array of oligos 3, each oligo being in a separate area 2 of support 1. Figure 1a shows one area 2. A binary oligo 3 (many copies, of course) comprised of constant region 5 and variable region 6 is cova-lently bound to support 1 by covalent linking moiety 4.

Because of the constant segments, binary arrays provide means for the hybridization of longer sequences without increas-ing the size of the array. The constant segment can be located within the immobilized oligo either "upstream" of the variable segment (i.e., toward or at the 5' end of the oligo) or "down-stream" from the variable segment (i.e., toward or at the 3' end of the oligo). The type of array that is chosen depends on the specific application. The constant region preferably is or includes a good priming region for amplification of hybridized strands by a polymerase chain reaction (PCR), or a promoter for copying the strand by transcription. Generally a length of 15 to 25 nucleotides is suitable for priming. The constant region can contain all or part of the complement of a restriction site. A binary array can be "plain" or "sectioned" (see below).

"Plain arrays" known in the art are arrays in which the individual areas are not physically separated from one another. Reactions carried out simultaneously are limited to those in which the nucleic acid templates and the reaction products are bound in some manner to the surface of the array to avoid the intermixing of products.

"Sectioned arrays" are divided into sections, so that each area is physically separated by mechanical or other means (e.g.,

-7-

a gel) from all the other areas, e.g., depressions on the surf-
ace, called a "well". There are many techniques apparent to one
skilled in the art for preventing the exchange of materials
between areas; any such method can be used to make a "sectioned"
array, as that term is used herein, even though there might not
be a physical wall between areas.

One type of sectioned array is illustrated in Figures 2 and
2a. Figure 2 shows a support sheet 60 having an array of depres-
sions or wells 62, each containing many copies of an immobilized
oligo 64. Figure 2a shows one well 62 of the array of Figure 2.
Well 62 formed in support 60 has therein oligo 64 covalently
bound to support 60 by covalent linking moiety 66. In practice
one may prepare a plain array, e.g., on a flat sheet, and then,
at a point during a series of steps involving its use, convert
the array into a sectioned array, e.g., by making physical
depressions in a deformable solid support to isolate the
individual areas. The sectioned array can also be created by
applying a lattice to the solid support and bonding it to the
surface so that each area is surrounded by impermeable walls. An
exploded perspective view of such a sectioned array is shown in
Figure 3. Support or substrate 70, here a planar sheet, has
mounted thereon and affixed thereto a lattice 72 comprised of a
series of horizontal members 74, 76. The lattice members define
a series of open areas which, in conjunction with support 70,
define an array of wells 78. In some applications it is prefer-
able to utilize a detachable lattice (or a removable cover
sheet), so that the sectioned array can be converted back to a
plain array.

Sectioned arrays according to this invention can be used to
increase the specificity of hybridization of nucleic acids to the
immobilized oligos. After hybridization, unhybridized strands
can be washed away. Hybridized strands can then be released into
solution without mixing. Released strands can be rebound to the
immobilized oligos, and unhybridized strands can be washed away.
Each successive release, rebinding, and washing increases the
ratio of perfectly matched hybrids to mismatched hybrids.

-8-

An array can be "3'" or "5'". "3' arrays" possess free 3'
termini and "5' arrays" possess free 5' termini. The immobilized
oligos in a 3' array can be extended at their 3' termini by
incubation with a nucleic acid polymerase. If it is a template-
directed polymerase, only immobilized oligos hybridized to a
template strand can be extended.

Methods of oligodeoxyribonucleotide synthesis directly on a
solid support are also known in the art, including methods
wherein synthesis occurs in the 3' to 5' direction (so that the
oligos will possess free 5' termini). Methods wherein synthesis
occurs in the 5' to 3' direction (so that the oligos will possess
free 3' termini) are also known.

Suitable substrates or supports for arrays should be non-
reactive with reagents to be used in processing, washable under
stringent conditions, not interfere with hybridization and not be
subject to inordinate non-specific binding. For example, treated
glass polymers of various kinds (e.g., polyamide and polyacromor-
pholide), latex-coated substrates and silica chips.

Arrays can be made over a wide range of sizes. In the
example of a square sheet, the length of a side can vary from a
few millimeters to several meters.

II. Sorting nucleic acids

Our invention allows mixtures of strands to be sorted
according either to their terminal oligo segments ("terminal
sorting") or their internal oligo segments ("internal sorting")
on a binary array.

There are two important aspects of our invention for sort-
ing. First, each strand in a mixture can be made to hybridize at
only a few, or a single, location. And second, each strand can
be provided with universal terminal priming regions that enable
PCR amplication without prior knowledge of the terminal nucleo-
tide sequences and without the need to synthesize individual
primers.

For terminal sorting, the priming region(s) can be made
essentially dissimilar from the sequences occurring in the

-9-

nucleic acids that are present in the mixture to be sorted, so
that priming does not occur anywhere but at the strands' termini.
When strands from a complete restriction digest of a DNA are to
be terminally sorted and amplified, priming only at the strands
termini can be promoted by restoring the terminal restriction
sites (those sites having been eliminated from internal regions
by complete digestion) concomitant with the generation of
terminal priming regions.

Terminal sorting is carried out on a binary array, which
preferably is sectioned. The immobilized oligos contain a
constant segment complementary to either the strands' 3' priming
region or 5' priming region. Thus, each strand can only be
hybridized to one location within the array. By sorting on a
comprehensive array, every strand is bound somewhere within the
array. This is especially important for the preparation of a
comprehensive library of fragments of a long nucleic acid or a
genome.

Strands can be sorted on either 3' or 5' arrays in which the
constant segment is located either upstream or downstream of the
variable segment. High specificity of sorting can be achieved by
employing 3' arrays in which the constant segment of the immobi-
lized oligos is upstream. In that case, sorting can be followed
by the generation of an immobilized copy of each sorted strand
using the immobilized oligos as primers for the synthesis of a
complementary copy of that strand when the array is incubated
with an appropriate DNA polymerase. The generation of copies
covalently linked to the array enables the array to be vigorously
washed to remove non-covalently bound material before strand
amplification. It also enables the arrays to serve as permanent
banks of sorted strands which can subsequently be amplified over
and over to generate copies for further use.

A strand sorting procedure is shown in Figure 4. A DNA
sample 10 is completely digested with a restriction endonuclease.
The ends of each fragment are restored, and universal priming
sequences 17 generated in the process to prepare fragments 11 for
sorting. It is not necessary that priming sequences be added at
both ends, if only linear amplification is desired. Nor is it

-10-

necessary that the priming sequence at the 3' end of a strand be
the same as the priming sequence at the 5' end.

The strands are then melted apart 12 and hybridized to a
terminal sequence binary sorting array, whose immobilized oligos
14 contain a variable segment 15 and a constant segment 16 which
is complementary to the universal priming region 17, including
the restored recognition site of the restriction enzyme 16a, 17a.
Each strand is at a location dependent upon its variable sequence
100 adjacent to its priming sequence. At this point the array
need not be sectioned. The array is then washed to remove
unhybridized strands. The entire array is then incubated with
DNA polymerase. Consequently, a complementary copy 18 of each
hybridized DNA strand is generated by extension of the 3' end of
the oligo to which the strand is bound. The array is then
vigorously washed to remove the original DNA strands and all
other material not covalently bound to the surface (not shown).

The covalently bound copy strands can be amplified. During
amplification it is usually desirable that the array be sec-
tioned. The wells are filled with a solution containing univer-
sal primers 19, 20, an appropriate DNA polymerase, and the
substrates and buffer needed to carry out PCR. The array can, if
desired, be sealed with a coversheet, further isolating the wells
from each other. PCR is carried out simultaneously in each well
of the array. This results in sorting the mixture of strands
into groups of strands that share the same terminal oligo
sequence, each strand (or each group of strands) being present in
a different well of the array and amplified there.

The results of hybridization can be improved by "proof-
reading", or editing, the hybrids formed, by selectively destroy-
ing those hybrids that contain mismatches, without affecting
perfect hybrids.

The length of the immobilized oligos in a strand sorting
array is chosen to suit the number of strands to be sorted. When
sorting strands according to their terminal sequences, the number
of different strands obtained in each well equals the number of
times that a particular oligo complementary to the variable
segment of the immobilized oligo occurs among the termini of

-11-

different strands in the mixture. If the number of nucleotides
in each variable segment is n, then the total number of such
variable sequences is 4^n, and the mean number of different
strands in a well is $N/4^n$, where N is the number of different
strands in the mixture, provided that nucleotide sequence is
random, and that each of the four nucleotides is present in equal
proportion. If a random sequence that is the size of an entire
diploid human genome (6×10^9 basepairs) is completely digested
by a restriction endonuclease that has a hexameric recognition
site, then the resulting mixture will contain approximately
3×10^6 strands with an average length of 4,096 nucleotides. If
this mixture is then applied to a comprehensive binary array
having variable segments eight nucleotides long, then each well
will contain, on average, approximately 45 different strands.

Our invention also includes methods for isolating individual
strands by sorting them according to the identity of their
terminal sequences on sectioned binary arrays. The strands can
be from restriction fragments or not, so long as unique priming
sequences are added to at least one of the strand's termini, such
as by methods described herein. If the number of different
strands in a sample is rather small, there is a high probability
that after the first stage of sorting, many wells will either not
be occupied, or be occupied by only one type of fragment. In the
case of a complex mixture of strands (such as from the digestion
of an entire human genome), a number of different types of
fragments will occupy each well. In that case, the isolation of
individual fragments can be achieved by PCR amplifying the
strands in each well in the first stage of sorting and then
sorting the group of fragments from each well on a fresh sec-
tioned array. After symmetric PCR amplification, each well of
the first array will contain copies of the strands that were
originally hybridized there, and also their complementary copies.

If the original strands were sorted by their 3' ends, then
their copies in a given well will all possess the same
3'-terminal sequence, and their complementary copies will possess
the same 5' end. However, the 3'-terminal sequences of the
complementary copies of the original strands in each well will be

-12-

different (as will be the 5' terminal sequences of the original
copies). Therefore, the complementary strands will bind at
different locations within the new sectioned array, according to
the identity of their own 3'-terminal sequences, and with a high
probability, each of them will occupy a separate well, where they
can then be amplified.

Alternatively, the second stage of sorting can be carried
out according to the identity of the terminal sequences at the
other end of each strand. For example, if the strands were
sorted in the first stage by their 3' ends (on an array whose
immobilized oligos contain upstream constant segments, then the
groups of strands from each well in the first array can be sorted
in a second stage by their 5' termini (on an array having down-
stream constant segments). In either procedure, as a result of
the second round of sorting, almost all of the different types of
fragments are separated from one another (with the exception of
virtually identical allelic strands from a diploid genome, which
usually have identical termini, and consequently are sorted into
the same well). The isolated strands can then be used for any
purpose. For example, they can be inserted into vectors and
cloned, or they can be amplified and their sequences determined.

Our invention also includes the use of binary arrays for
isolating selected strands by sorting according to the identity
of terminal sequences. Strands can, for example, be selected
that contain particular regions (such as genes) of special
interest from a clinical viewpoint. After the relevant portion
of a genome has been sequenced, an array can be made using only
preselected oligos whose variable segments uniquely match the
terminal sequences of the strands of interest, i.e., they would
be long enough to uniquely hybridize to the desired strands.

Our invention also encompasses methods that include sorting
fragments according to their internal sequences. When so sort-
ing, strands may bind at more than one well. This type of
sorting can be useful for a number of applications, such as the
isolation of strands that contain particular internal sequence
segments (utilizing a sectioned ordinary array), or the sorting
of strands according to the identity of variable oligo segments

-13-

adjacent to internal restriction sites of a particular type
(utilizing a sectioned binary array). The latter approach is
useful for ordering sequenced restriction fragments. The sorting
of strands by their internal segments on a 3' sectioned ordinary
array is useful for the generation of partial strands by virtue
of extension of the immobilized oligos.

Our invention includes the sorting, in particular for
sequencing, of natural mixtures of RNA molecules, such as
cellular RNAs. Establishing messenger RNA sequences is useful,
not only for the identification and localization of genes in the
genomic DNA, but also for providing information necessary to
determine the coding gene sequences (i.e. the exon/intron struc-
ture of each gene). Furthermore, the analysis of cellular RNAs
in different tissues, at different stages of development, and in
the course of a disease, will clarify which genes are active.
Usually, RNAs are short enough to be sorted and analyzed without
preliminary fragmentation.

III. Preparing partial strands of nucleic acids on sectioned
     arrays

Our invention includes methods of using sectioned arrays for
preparing all possible partial copies of a strand or a group of·
strands. Preparing complete sets of partials of a strand(s), and
sorting the partials by their variable ends is especially useful
in a process for determining the sequence of the strand or
strands. The preparation of partials is accomplished by either
of the following methods: (1) terminally sorting on sectioned
binary arrays a mixture of partial strands generated by degrada-
tion of a "parental" strand(s) at random; or (2) generating
partials on a sectioned ordinary array, through the sorting of a
parental strand(s) according to the identity of the strand's
internal sequences, followed by the synthesis of (complementary)
partial copies of the parental strand(s) by the enzymatic exten-
sion of the immobilized oligos, utilizing the hybridized parental
strands as templates, and then copying the immobilized partials.

-14-

By using comprehensive arrays, it is possible to prepare every
possible one-sided partial of a strand.

In the first case (partialing before sorting), a strand, or
a double-stranded fragment, or a group of either, carrying
terminal priming regions, (these can be a strand or a group of
strands sorted on a sectioned binary array as described above),
is randomly degraded by a chemical or an enzymatic method, or by
a combination of both. Then the mixture of partials is sorted on
a sectioned binary array according to the identity of their newly
generated termini, essentially as described above for the sorting
of full-length strands by their terminal sequences, with new
priming sites being introduced at these new termini either before
or after sorting. Only those partials that possess both the
newly introduced priming site and the already existing priming
site (at the opposite end), will be amplified by subsequent PCR.
Partials can be sorted according to the identity of a variable
sequence at either their 3' termini or their 5' termini.
However, as is the case for the sorting of full-length strands,
the highest specificity can be achieved by sorting according to
the identity of a variable sequence at the 3' termini, and
carrying out the sorting on 3' arrays having upstream constant
segments, or by sorting according to the identity of a variable
sequence at the 5' termini, and carrying out the sorting on 5'
arrays having downstream constant segments. In these cases,
sorting can be followed by the generation of immobilized
(complementary) copies of the sorted partials. The arrays with
the immobilized copies can serve as permanent banks of the sorted
partials which can subsequently be amplified over and over to
generate copies for further use. Following sorting, each well in
the array will contain immobilized copies of all of those
partials whose variable end is complementary to the variable
segment of the immobilized oligo. The other (fixed) end of these
partials will be identical to one of the ends of the parental
strands. If an oligo segment occurs more than once in a strand,
or if it occurs in more than one strand in the group of strands
subjected to partialing, then the well will contain a

-15-

corresponding number of different partials, all sharing the same sequence at their variable ends.

In the second case (sorting before partialing), partials are prepared directly from the parental strands that are hybridized to a sectioned ordinary array without prior degradation. A strand, or a mixture of strands, is hybridized to a 3' ordinary array. The immobilized oligos are then used as primers for copying the hybridized strands, beginning at the location within each bound strand where hybridization occurred, and ending at the upstream terminus of each bound strand. After extension of the immobilized oligos, the hybridized parental strands are discarded. At this point the wells contain immobilized (complementary) partial strands. The partials in one well all share a 5'-terminal oligo segment that is complementary to a particular internal oligo in the parental strand(s). The partial strands have 3'-terminal sequences that include the complement of the 5'-terminal region of the parental strand(s) (which contains a priming region). Unlike the methods described above for partialing before sorting, the immobilized complementary partials will contain a priming region at only one end and therefore can not be amplified exponentially. However, their linear amplification is possible, with the partials being synthesized as DNAs or RNAs. Where RNA partials are generated, the priming region at the partial copy's 3' terminus contains an RNA polymerase promoter. Synthesis of RNA copies is more efficient than linear synthesis of DNA copies. Alternatively, the synthesized copies can be provided with second priming regions and can then be amplified in an exponential manner by PCR. This approach is illustrated, schematically, in Figure 5.

Figure 5 illustrates the generation of partials for one DNA parental strand 30 on a 3' sectioned ordinary array. First, the strand 30 (many copies, of course) such as obtained from well 13a of sorting array 13, is hybridized to the partialing array 31, a 3' sectioned ordinary array, containing well 31a. The parental strand 30 binds to many different locations within the array, dependent on which oligo segments are present in the strand. A hybrid 32 is formed in each well at the array that contains an

-16-

immobilized oligo complementary to a strand's oligo segment.
After hybridization, the entire array is washed and incubated
with an appropriate DNA polymerase in order to extend the im-
mobilized oligos utilizing the hybridized strand as a template.
Each extension product 33 strand is a partial (complementary)
copy of the parental strand. Each partial begins at the place 32
in the strand where hybridization occurred and ends at the
strand's terminus. The strand preferably terminates at its 5'
terminus with a universal priming sequence 17, such as one
introduced into all strands when sorting strands on a sectioned
binary array as described. This allows for amplification of the
partials. That priming sequence can contain a restored restric-
tion site 16a. The parental strand may also contain, if it was
previously sorted on a binary sorting array, a priming sequence
at its 3' terminus 17, adjacent to the variable sequence 100 that
the strand was previously sorted by.

     The entire array is then vigorously washed under conditions
that remove the parental DNA strands and other material, prefer-
ably all, that is not covalently bound to the surface. The areas
of the array then contain immobilized strands 33 that are com-
plementary to a portion of the parental strand. The wells can
then be filled with a solution containing the universal primer
(or promoter complement), an appropriate polymerase, and the
substrates and buffer needed to carry out multiple rounds of
copying of the immobilized partial strands. The array can then
be sealed, isolating the wells from each other, and (linear)
copying can be carried out simultaneously in all of the wells in
the array.

     IV. Surveying oligonucleotides with binary arrays

     Our invention includes using binary arrays to survey oligos
contained in strands and partials. Binary arrays allow surveying
to be improved as compared with ordinary arrays, and they allow
new types of selective surveying (such as surveying "signature
oligonucleotides").

-17-

In surveying, strands first can be randomly degraded into
pieces whose average length slightly exceeds the surveyed length.
After degradation, each resulting nucleic acid piece is ligated
to the same type of oligo (i.e., a constant sequence), that
preferably does not occur anywhere in the internal regions of the
pieces. For example, the sequence of the added oligo can contain
the recognition site of a restriction endonuclease that was used
to digest the DNA prior to fragment sorting. The ligation can be
carried out in solution prior to hybridization, or after hybridi-
zation of the pieces to binary immobilized oligos whose constant
segment is complementary to the oligo to be ligated. Preferably,
a 3' array is used, having upstream constant segments. The
immobilized oligos can then be extended with an appropriate DNA
polymerase, using the hybridized nucleic acid pieces as
templates. It is preferable that after extension all hybrids
have the same length. This can be achieved by employing dideoxy-
nucleotides as substrates for the polymerase, to restrict exten-
sion to one nucleotide.

Hybrids can be labeled in both a ligation-dependent and an
extension-dependent manner to increase the specificity of hybrid
detection. Also, the ligated oligos and the added dideoxy-
nucleotides can be tagged with different labels, for example,
fluorescent dyes of different colors. The array is then scanned
at two different wavelengths, and only those areas that emit
fluorescence of both colors indicate perfect hybrids.

Survey results can be improved further by hybrid proof-
reading, by destroying hybrids containing mismatches, and by
using chemical or enzymatic methods.

V.   Use of the oligonucleotide arrays for the sequencing of
      nucleic acids

The arrays and methods of this invention can be used to
determine the nucleotide sequence of nucleic acids, including the
sequence of an entire genome, whether it is haploid or diploid.
This embodiment requires neither cloning of fragments nor prelim-
inary mapping of chromosomes. It is especially significant that

-18-

our method avoids cloning, a labor-intensive and time-consuming
approach that is essentially a random search for fragments. In a
preferred embodiment a comprehensive collection of whole nucleic
acids or fragments is sorted into discrete groups. The sorted
nucleic acids are then amplified with a polymerase, preferably by
PCR.

Sequencing large diploid genomes, such as a human genome,
using the arrays and methods of this invention is shown in Figure
6. We will describe the overall method in general terms. In the
embodiment illustrated in Figure 6 an individual's genomic DNA 40
is digested with a restriction endonuclease and sorted by ter-
minal sequences into groups of strands using a 3' sectioned
binary sorting array 13, as is described above in Section II and
illustrated in Figure 4.

Next, treating each well 13a of the sorting array separ-
ately, a complete set of partials is prepared for each group of
sorted strands using a sectioned array 31, as is described above
in Section III and illustrated in Figure 5. The partials can be
generated in any chosen manner to make them detectable.

Then the contents of each well 31a of the partialing array
31 is surveyed using a survey array 42, as is described above in
Section IV. Preferably the survey array is a binary array, but
an ordinary array may be used. In the embodiment shown in Figure
6, surveying is performed with a sheet 43 containing miniature
survey arrays 42 that have been printed in a pattern that coin-
cides with the number and location of the wells 31a. The oligo
information obtained can be used, according to our invention, to
separately determine the nucleotide sequence of every strand in
each group isolated on the sorting array.

To determine the order of the fragments sequenced as il-
lustrated in the embodiment of Figure 6, genomic DNA 40 is
digested with at least a second restriction endonuclease and
sorted into groups of strands using a 3' sectioned binary sorting
array 44, as is described above in Section II and illustrated in
Figure 4. The contents of each well 44a of the sorting array 44
is surveyed with special survey arrays 45, 46 that identify
"signature oligonucleotides" (described below) in intersite

-19-

segments of sorted fragments from different digests. This is
done to determine the order of the fragments relative to one
another without regard to differences between allelic pairs of
fragments. In the embodiment shown in Figure 6 this surveying is
performed with printed sheets 47, 48 that have been printed with
a pattern of miniature arrays 45, 46.

To allocate the ordered allelic fragments to their respec-
tive chromosomes in a diploid organism, fragments are linked
according to their allelic differences. In the embodiment
illustrated in Figure 6, the strands from selected wells of the
sorting array 44 are transferred to a selected well of one of a
series of partialing arrays 49, partials are generated, and the
partials are surveyed using miniature survey arrays 50 on printed
sheets 51. Only the presence of oligos containing allelic
differences in the selected partials needs to be determined to
link a pair of allelic fragments to their respective neighboring
allelic fragments.

When sorting according to the identity of terminal sequen-
ces, each strand occupies a particular "address" in the array.
It is convenient to think of the address as the oligo sequence
within a strand that directs the DNA strand to hybridize to a
particular location, i.e., the sequence that is perfectly com-
plementary to the variable sequence of the oligo immobilized at
that location. The "address" also identifies the location within
the array where the DNA binds.

After sorting, each group of strands is amplified and
subjected to partialing. Importantly, the isolation of
individual strands is not necessary, because our method allows
the nucleotide sequence of each strand in a mixture to be deter-
mined. In particular, our method allows the sequences of strands
in a well of the sorting array to be determined, separately from
mixtures of strands in other wells. In a preferred embodiment,
the partialing array is comprehensive in order to obtain all
possible one-sided partials (i.e., a comprehensive array). Each
group of partials is amplified prior to surveying. Most preferab-
ly, the amplification is carried out in such a manner that one

-20-

of the two complementary partial strands is produced in great
excess over the other.

Each group of partials is surveyed to identify their con-
stituent oligos. Surveying is preferably carried out using
binary arrays.

Although not necessary, it is preferable to have the survey
arrays be as compact as possible. It is anticipated that survey-
ing will be advantageously accomplished simultaneously for many
or all wells of a partialing array by utilizing a sheet on which
miniature survey arrays have been "printed" in a pattern that
coincides with the arrangement of wells in the partialing array,
in a manner similar to that shown in Figures 6 and 7. Referring
to Figure 7, partialing array 31, comprising an array of wells
31a, is surveyed using sheet 43, having printed thereon an array
of miniaturized survey arrays 42. The pattern of arrays 42
corresponds to the pattern of wells 31a, whereby all wells 31a
can be surveyed simultaneously.

Automated photolithography techniques for preparing minia-
ture oligo arrays have been developed [Fodor, S. P., Read, J. L.,
Pirrung, M. C., Stryer, L., Lu, A. T. and Solas, D. (1991).
Light-Directed, Spatially Addressable Parallel Chemical
Synthesis, Science 251, 767-773]. The manufacture of miniature
arrays on a "chip", for use in surveys also has been reported.

Surveying with comprehensive arrays produces a complete list
of oligos contained in the partials in each well of the partial-
ing array. This will reveal all oligos present in all partials
in that well. The method of this invention can determine the
sequences of the original (parental) fragment strands.

The "partials" referred to in this section are one-sided
partial strands that begin at the 5' terminus of a parental
nucleic acid strand (the fixed end) and end at different nucleo-
tide positions in the strand (the variable end). Partials are
sorted in the partialing array according to the identity of their
variable ends, and therefore each partial has a particular
"address" within the array. As with sorting arrays, an "address"
in a partialing array is the oligo sequence that is present at
the variable end of the partial strand and that is complementary

-21-

to the variable segment of an immobilized oligo. The "address"
also relates to the location within the array where the partial
strand is found, since the variable segment of the oligo immobi-
lized in that well is complementary to the oligo at the partial's
variable terminus. The "address" also relates to the location
within the parental strand of a partial's terminal oligo. The
location of this "address oligo" within a parental strand is
characterized by an "upstream subset" of oligos that come before
it in the parental sequence and by a "downstream subset" of
oligos that come after it.

Our method of establishing nucleic acid sequences, for
either a single strand or a group of parental strands sorted by
their terminal sequences, begins by assembling an "address set"
for each address in the partialing array. The "address set" is a
comprehensive list of all oligos in all the parental strands
which have the address oligo within their nucleotide sequences.
The "upstream subset" contains all the oligos that occur upstream
(i.e., towards the 5' end) of the address oligo in parental
strands that contain the address oligo. The "downstream subset"
contains all the oligos that occur downstream (i.e., towards the
3' end) of the address oligo in any parental strands that contain
the address oligo. Together the two subsets form the "address
set."

The upstream subset of each address can be determined
directly from the survey of each well of a partialing array and
consists of a list of all the oligos identified as being present
in the partial strands in that well. The downstream subset of
each address can be inferred by examining the upstream subsets of
all the addresses: the downstream subset of a particular address
consists of those addresses whose own upstream subset includes
that particular address oligo.

The upstream subset and the downstream subset of a par-
ticular address, taken together, are an "indexed address set".
If an oligo occurs more than once in a strand, it can occur in
both the upstream and the downstream subsets of an address.
Indexed address sets provide the information required to order
the oligos contained in a strand set, as will be described below.

-22-

When a mixture of strands is examined, it is also useful to
consider an address set without regard to which oligos occur
upstream and downstream of an address. This is called an
"unindexed address set". Unindexed address sets are decomposable
into strand sets by the method of this invention.

We have discovered that when assembling big strand sets
whose oligos do not all overlap uniquely, it is advantageous to
work with "sequence blocks" rather than with individual oligos.
Sequence blocks are composed of oligos that uniquely overlap one
another in a given strand set. Two oligos contained in a strand
set are said to overlap if they share a terminal (5' or 3') $n-1$
nucleotide sequence. An overlap is unique if no other oligo than
those two in the strand set has this sequence at its termini.
Here n is the length (in nucleotides) of each of the two oligos
if they are of the same length or, if they are of different
length, n is the length of the shorter one. We use unique
overlaps to construct sequence blocks from the oligos in a strand
set.

The position of each sequence block relative to the others
is determined from the distribution of the oligos between the
upstream and downstream subsets of every address. This is
accomplished by finding, for each of the blocks, which blocks
occur upstream, and which blocks occur downstream, of that block
by examining the address sets. The address sets are used in
order to generate "block sets." The block sets are address sets
wherein blocks have been substituted for the oligos that comprise
the blocks, including the address oligo. Once the relative
position of the sequence blocks has been determined, they can be
assembled into the final sequence. The assembly is governed by
the following rules: (1) each of the blocks must be used at
least once, (2) the blocks must be assembled into a single
sequence, (3) the ends of neighboring blocks must match each
other (i.e., overlap by an $n-1$ nucleotide sequence, see above)
and (4) the order of the blocks must be consistent with their
positions relative to one another, as ascertained from the block
sets, as will be clear from the examples.

-23-

A sequence block can occur either once in a sequence, or
more than once, and this we determine by examining the block
sets. If a block occurs more than once in a sequence, it will
always be contained in both its own upstream and downstream
subsets. On the other hand, if a block occurs only once in a
sequence, it may or may not be present in its own upstream or
downstream subset. But, if a block is absent from either its
upstream subset, or from its downstream set, that block occurs in
the strand only once. The relative order of these "unique"
blocks can be determined by noting which of them occur in the
upstream subset, and which of them occur in the downstream
subset, of the others. Once the unique blocks have been ordered
relative to each other, the gaps between them are filled with
blocks that may be non-unique. However, not every gap can
necessarily be filled in with a particular block. There is a
range of locations within which each non-unique block (or
presumably-non-unique block) can be present. The range for a
particular block is determined by noting those blocks that always
occur upstream of it, and those blocks that always occur down-
stream of it. A gap can be filled in if, and only if, there is a
block or a combination of blocks, whose outer ends have $n-1$
nucleotide-long perfect sequence overlaps with the ends of the
blocks that form the gap. Because at least two overlaps, each of
low probability, must occur simultaneously, it is highly unlikely
that more than one block, or one combination of blocks, can fill
a gap. If a particular block occurs many times in a strand, it
will have to be used to fill every gap it matches. This is why,
using the method of the invention, it is possible to establish
the sequence of a strand without measuring how many times an
oligo occurs in the partials. It is only necessary to determine
whether an oligo is present or not.

An important aspect of this invention is the ability to
sequence a mixture of strands simultaneously. The invention can
be used for the determination of fragment sequences from an
entire fragmented and sorted genome.

If one strand is being sequenced, all address sets deter-
mined from a partialing array will contain the same oligos that

-24-

constitute the strand set. The only difference is that some
oligos which are downstream in one set may be upstream in another
address set. If a mixture of strands have been partialed on a
single partialing array, certain addresses will be shared by more
than one parental strand. Their address sets will be composite,
containing all of the oligos from all of the strands that the
address oligo is present in. Addresses that are only found in a
particular strand in the mixture, however, will have address sets
which only contain oligos from that strand. They are identical
to the strand set, and each contain the same oligos. The mixture
can contain up to a hundred or so different DNA strands, each of
a different length and sequence, as can be obtained with an
appropriate sorting array (or set of sorting arrays) and method
described above. When a mixture of strands is analyzed on a
partialing array, the data obtained by surveying the partials
will reflect the diversity of the sequences in the mixture, and
will appear to be very complex. However, we have discovered a
way to decompose the unindexed address sets obtained by analysis
of a strand mixture into their constituent strand sets. Then, as
we have described for sequencing a single strand, the oligos in
each of the identified strand sets can be grouped into sequence
blocks that can be ordered from the information contained in the
indexed address sets, as will be clear from the examples.

     Unindexed address sets can be either "prime" or "composite."
A prime set consists of one strand set; while a composite set
consists of more than one. A prime set cannot be decomposed into
other address sets, i.e., there is no address set which is a
subset of a prime set. Composite sets, however, can usually be
decomposed into two or more simpler address sets. Once
individual strand sets have been identified, they can each be
treated as though they were obtained from an analysis of a
homogeneous strand. It is thus possible, in many cases, to
sequence all strands in an unknown heterogenous DNA sample
without first isolating the strands.

     The fragment sequences obtained by the methods outlined
above or by any other method can then be put in their correct
order using oligo arrays. Assembling restriction fragments into

-25-

contiguous sequences can be accomplished by identifying each
fragment's immediate neighbors. One method for obtaining this
information is to use another restriction enzyme to cleave the
same DNA at different positions, thus producing a set of frag-
ments that partially overlap neighboring fragments from the first
digest, and then to sequence these fragments. However, it is not
necessary to sequence the fragments in the second restriction
digest. It is only necessary to uniquely identify overlapping
segments in the fragments from alternate restriction digests.
This can be done by surveying "signatures".

Signatures can be determined by hybridization of fragment
strands to complementary oligo probes. A signature of a fragment
may consist of one, two or more oligos, so long as it is unique
within the sequence analyzed. Neighboring fragments from one
restriction digest can be determined by looking for their signa-
tures in overlapping fragments from an alternate digest.

We have devised a method for identifying neighboring
restriction fragments among the list of sequenced fragments that
does not require either cloning or sequencing of overlapping
fragments. If strands from an alternate digest are sorted,
complementary strands of the same fragment will hybridize to
different addresses in the sorting array. Whenever intersite
segments from two or more fragments of the first digest are
present within one fragment of the second digest, then all of
these segments will be represented in both complementary strands
of that one fragment, and all will be present wherever those
strands bind in a sorting array. We identify the segments by
obtaining their signatures through hybridization to specialized
binary survey arrays. The signatures of intersite segments that
occur in one fragment always accompany each other, whereas
signatures of distant segments travel independently.

After the fragments from an original (first) restriction
digest of a long DNA have been sequenced, the same DNA is
digested with a second (different) restriction endonuclease, the
termini of the generated fragments are provided with universal
priming regions (that also restore the recognition sites at the
termini), and the strands are sorted according to particular

-26-

internal sequences, namely, a variable sequence adjacent to the recognition site for the first restriction enzyme. The sorting array is a sectioned binary array. It contains immobilized oligos having a variable sequence as well as an adjacent constant sequence that is complementary to the recognition sequence of the first restriction endonuclease. The sorted strands are amplified by "symmetric" PCR, so that in each well where a strand has been bound, copies of the bound strand, as well as complements, are generated. In another embodiment, strands can be sorted according to their terminal sequences on an array whose oligos' constant segments include sequences that are complementary to the recognition site of the second restriction enzyme. This alternative is not detailed, but it corresponds to the embodiment discussed below, but with terminal sorting.

Each strand that hybridizes to the binary sorting array will possess at least two recognition sites for the second restriction enzyme (restored at the strand's termini), and at least one (internal) recognition site for the first restriction enzyme. The segments included between these two types of restriction sites (intersite segments) comprise the overlaps between the two types of restriction fragments, and each intersite segment is thus bounded by any two restriction sites of the two types. It follows, that each of these segments can be characterized by identifying these two restriction sites and variable sequences of preselected length within the segment that are immediately adjacent to each of the restriction sites. The combination of a recognition site (for either the first or the second restriction enzyme) and its adjacent variable oligo we call a "signature oligonucleotide". Every intersite segment can be characterized by two signature oligos (of either type) that bound that segment. The combination of the two signature oligos is defined herein as the intersite segment's "signature".

After strand amplification, the strands in the wells of the sorting array are surveyed to identify the signature oligos of each of the two types. This is carried out by using two types of binary survey arrays. The first has immobilized oligos containing a variable oligo segment and a constant segment that is, or

-27-

includes, an adjacent sequence that is complementary to the
recognition site for the first restriction endonuclease. The
immobilized oligos in the second survey array has a variable
oligo segment of preferably the same length as the variable
segment of the first specialized survey array, and a constant
segment that is, or includes an adjacent sequence that is com-
plementary to the recognition site for the second restriction
endonuclease. The constant oligo segments in these arrays can be
located either upstream or downstream of the variable oligo
segments, resulting in the surveying of either the downstream or
the upstream signature oligos in each strand of the intersite
segments being surveyed. In a preferred embodiment the constant
oligo segments are upstream, and the immobilized oligos have free
3' ends, so that they can be extended by incubation with a DNA
polymerase. From the oligo information that is obtained, the
sequenced fragments can be ordered relative to one another.

       In our method, the uniqueness of a signature is achieved by
surveying "half signatures" (signature oligonucleotides) on two
relatively small survey arrays. If the variable segments in the
arrays are 8-nucleotide-long, the number of areas in the two
arrays is approximately 130,000, or approximately 100,000,000
times smaller than the single array that would be needed for
detecting the same size signature (28 nucleotides).

       If a diploid genome (such as a human genome) is sequenced,
the ordered fragments will appear as a string of unlinked pairs
of allelic fragments. What remains unknown is how the allelic
fragments in each pair are distributed between the homologous
(sister) chromosomes that came from each parent. Allocation of
the allelic fragments to these "chromosomal linkage groups"
requires knowledge of which fragment in each pair is linked to
which fragment in a neighboring pair.

       We have developed a method that uses arrays for allocating
allelic fragments to chromosomes, irrespective of what method was
used for sequencing and ordering the fragments. The linkage of
fragments in neighboring pairs can be achieved by sequencing a
restriction fragment ("spanning fragment") from an alternate
digest that spans at least one allelic difference in each pair.

-28-

Since the sequences of the allelic fragments are known, there is
no need to sequence the spanning fragment. Instead, one can
simply determine which oligos that harbor allelic differences
accompany one another in the spanning fragment, i.e., which
oligos occur in the same chromosome. This can be accomplished by
surveying, at a selected address in a partialing array, partials
generated from a selected group of restriction fragments from an
alternate digest. A group of restriction fragments is selected
that contains a spanning fragment, and an address in a partialing
array is selected that encompasses a difference in one of the
neighboring allelic pairs.

Since the sequence of every fragment is known, it is pos-
sible to choose an alternate restriction fragment that spans the
allelic differences in the neighboring pairs. A spanning re-
striction fragment, in fact, may already be present at a par-
ticular address in one of the sorting arrays used to sort alter-
nate digests during the ordering procedure.

In this method, sorted strands are melted apart, and the
mixture is hybridized to a particular well in the partialing
array, whose address corresponds to one of the allelic oligos.
Two different wells are selected, each with an address that
corresponds to an oligo that harbors a differenct allelic oligo-
nucleotide After amplification of the partial strands, the oligos
in the two wells are identified with a survey array. Examination
tells which fragments are on the same chromosome.

Since allelic differences occur roughly once every 1,000
basepairs in the human genome, most allelic fragments resulting
from digestion with a restriction enzyme recognizing a hexameric
sequence (resulting in about 4,096 average length) will differ
from each other. If the variable oligo segments in the survey
arrays are made of octanucleotides, then each allelic nucleotide
substitution will give rise to eight different oligos in each of
the allelic fragments. However, using our method, inspection of
only one address in the partialing array is sufficient to reveal
the linkage of the corresponding reference oligo to any one of
the eight oligos that encompass the nucleotide substitution that
occurs in the neighboring fragment on the same chromosome.

-29-

Therefore, only one address in the partialing array is needed to
reveal the linkages between two neighboring allelic pairs. Thus,
65,536 linkages can be determined on a single comprehensive
partialing array made of variable octanucleotides. With this
method, only 10 to 20 of these arrays would be needed to complete
the assembly of an entire diploid human genome that has been
fragmented by a restriction endonuclease with a hexameric recog-
nition site.

Computational methods can be developed to minimize or
eliminate errors that occur during partialing and surveying, by
taking advantage of the high redundancy in the data. Such
methods should take into account the following aspects of a
preferred sequencing procedure: the sequence of every fragment
is independently determined four times (by virtue of each strand
and its complement being present at two different addresses in
the sorting array); each strand set is determined in as many
trials as the number of different oligos in that strand; every
nucleotide in a strand is represented by as many different oligos
as the length (of the variable segment) of the immobilized oligos
in the survey array; the locations where a particular block can
occur in a sequence are limited by the distribution of the blocks
among the upstream and downstream subsets of each pertinent
address; and the edges of a block must be compatible with the
edges of each gap where that block is inserted.

Using our genome sequencing method, one can use throughout
essentially the same technology, i.e., hybridization of oligo
probes and the amplification of nucleic acids by the polymerase
chain reaction, both of which are well-studied, common laboratory
techniques. The entire procedure can be performed by a specially
designed machine, resulting in huge reductions in time and cost,
and a marked improvement in the reliability of the data. Many
arrays could be processed simultaneously on such a machine. The
machine most preferably should be entirely computer-controlled,
and the computer should constantly analyze intermediate results.
As stated above, used arrays can be stored, both to serve as a
permanent record of the results, and to provide additional

-30-

material for subsequent analysis or for manipulating the
sequenced strands and partials.

Analysis of an individual's genomic DNA provides the com-
plete nucleotide sequence of that individual's diploid genome.
The genes and their control elements are allocated into chromo-
somal linkage groups as they appear in a single living organism.
The sequence will describe an intact, functioning ensemble of
genetic elements. This complete sequencing provides the ability
to compare genomes of individuals, thereby enabling biologists to
understand how genes function together and to determine the basis
of health and disease. The genomes of any species, whether
haploid or diploid, can be sequenced.

The invention can be used not only for DNA's but as well for
sequencing mixtures of cellular RNAs.

The invention is also useful to determine sequences in a
clinical setting, such as for diagnosis of genetic conditions.

VI. Manipulating Nucleic Acids on Sectioned Arrays

Our invention also includes using sectioned arrays for
introducing site-directed mutations into sequenced nucleic acids,
including the introduction of nucleotide substitutions, deletions
and insertions. This can be carried out in a massively parallel
fashion. In one embodiment, a partial whose variable end has
been deprived of a priming region, is ligated to the free ter-
minus of an immobilized oligo that contains the mutation to be
introduced. In another procedure, where the purpose of muta-
genesis is to introduce a single-nucleotide substitution, then
the substituting nucleotide can be added directly to the variable
end of the partial. In both cases, the modified partials or
their complementary copies are used to synthesize a mutant strand
utilizing as a template either the complementary parental strand
(i.e., from which the partials were generated) or a longer
complementary partial, or any other strand or partial that
encodes the missing region. The fixed end of the mutant partial
is provided with a priming region that is different from the
corresponding priming region of the template strand. Therefore,
only mutant strands are capable of subsequent amplification by

-31-

PCR. A single array can be used either to mutate many single
positions in a gene, or to introduce mutations in many genes in
one procedure.

Sectioned arrays can also be used for the massively parallel
testing of the biological effects of the introduced mutations.
For example, parallel coupled transcription-translation reactions
can be carried out in the wells of a sectioned array following
amplification of the mutant strands. It is thus possible to
determine simultaneously, on the same sectioned array, the
effects of many different amino acid substitutions on the struc-
ture and function of a protein.

VII. Examples

1. Sorting nucleic acids or their fragments on a binary
oligonucleotide array whose immobilized oligos have free 3'
termini, with constant upstream segments --

This method allows the immobilized oligos to serve as
primers for copying bound strands, resulting in the formation of
complementary copies covalently linked to the array.

1.1. Sorting restriction fragments according to their
terminal sequences, following the introduction of terminal
priming regions --

DNA is digested using a restriction endonuclease. Recogni-
tion sites for the restriction endonuclease are restored in
solution by introducing terminal extensions (adaptors) that
contain a sequence which, together with the restored restriction
site, form a universal priming region at the 3' terminus of every
strand in the digest. This priming region is later used for
amplification by PCR. After melting fragments, the strands are
sorted on a sectioned binary array. A sequence complementary to
the generated priming region serves as both the constant segment
of the immobilized oligos and as the primer for PCR amplification
of the bound strands.

DNA to be analyzed is first digested substantially com-
pletely with a chosen restriction endonuclease, and the fragments

-32-

obtained are then ligated to synthetic double-stranded oligo
adaptors. The adaptors have one end that is compatible with the
fragment termini. The other end is not compatible with the
fragments' termini. The adaptors can therefore be ligated to the
fragments in only one orientation. The adaptors' strands are
non-phosphorylated, which prevents their self-ligation. The
strands in the restriction fragments have their 5' termini phos-
phorylated which results from their cleavage by a restriction
endonuclease. This favors the ligation of the adaptors by a DNA
ligase (such as the DNA ligase of T4 bacteriophage) to the
restriction fragments, rather then to each other. Since DNA
ligase catalyzes the formation of a phosphodiester bond between
adjacent 3' hydroxyl and phosphorylated 5' termini in a double-
stranded DNA, the phosphorylated 5' termini of the fragments are
ligated to the adaptor strand whose 3' end is at the compatible
side of the adaptor. The 3' termini of the fragments remain
unligated. A DNA polymerase possessing a 5'-3' exonuclease
activity (such as DNA polymerase I from *Escherichia coli* or Taq
DNA polymerase from *Thermus aquaticus*) is then used to extend the
3' ends of the fragments, utilizing the ligated oligo as a
template, concomitant with displacement of the unligated oligo.
To make the ligated oligo resistant to the 5'-3' exonuclease, the
ligated oligo can be synthesized from α-phosphorothioate precur-
sors.

Although the oligo adaptors are provided in great excess
during the ligation step, there is still a low probability that
two restriction fragments will ligate to one another, rather then
to the adaptor. To prevent this, the ligation products can again
be treated with the restriction endonuclease used to generate the
fragments, in order to cleave the formed interfragment dimers.
The endonuclease will not cleave the ligated adaptors if they are
synthesized from modified precursors (such as nucleotides con-
taining N^6-methyl-deoxyadenosine), which are known and currently
commercially available [e.g., from Pharmacia LKB]. Resistance of
the ligated adaptors to digestion by the restriction endonuclease
can be increased further if the ligated oligo is synthesized from
phosphorothioates, and if phosphorothioate analogs of the nucleo-

-33-

side triphosphates are used as substrates for extension of the 3'
termini.

After the priming regions have been added, the complementary
strands are melted apart, such as by increasing temperature
and/or by introducing denaturing agents such as guanidine iso-
thiocyanate, urea, or formamide. The resulting strands are
hybridized to a binary sorting array, such as by following a
standard protocol for the hybridization of DNA to immobilized
oligos. Hybridization is performed so that formation of only
perfectly matched hybrids is promoted. The hybrids have a length
which is equal to that of the immobilized oligos. The immobi-
lized oligos are attached to the array at their 5' termini and
contain constant restriction site segments adjacent to a variable
segment of predetermined length. Each strand will be bound to
the array at its 3' terminus. Its location within the array will
be determined by the identity of the oligo segment that is
located in the strand immediately upstream from the restored
restriction site at its 3' end, and that is complementary to the
variable segment of the immobilized oligo to which it is bound.
After hybridization and washing away all unbound material, the
entire array is incubated with a DNA polymerase, such as Taq DNA
polymerase deoxyribonucleotide 5' triphosphates or the DNA
polymerase of bacteriophage T7, and substrates. As a result, the
3' end of each immobilized oligo to which a strand is bound will
be extended to produce a complementary copy of the bound strand.
The array is vigorously washed. The wells are then filled with a
solution containing universal primer, an appropriate DNA polymer-
ase, and the substrates and buffer needed to carry out PCR. The
array is then sealed, isolating the wells from each other, and
exponential amplification is carried out, preferably simul-
taneously, in each well.

1.2. Sorting restriction fragments according to their
terminal sequences, with 3' and 5' terminal priming regions being
introduced, one before and one after strand sorting --
This procedure consumes larger amounts of enzymes and
substrates than the procedure described in Example 1.1, however,

-34-

only those strands that are correctly bound to the immobilized
oligos acquire both priming regions necessary for PCR. The
possibility that non-specifically bound strands will be amplified
is minimized. Furthermore, different priming regions can be
introduced at different termini of a strand. It then becomes
possible to: (1) perform "asymmetric" PCR, where only one of the
complementary strands is accumulated in significant amounts, and
remains single-stranded: (2) introduce a transcriptional promoter
into only one of the priming regions, in order to be able to
obtain RNA transcripts of only one strand (without also producing
its complement; (3) differentially label complementary strands;
and (4) avoid self-annealing of the strand's terminal segments
that can interfere with primer hybridization and lower PCR
efficiency.

     In this example, digestion of DNA, adaptor ligation and re-
digestion of fragments are carried out as described in Example
1.1, above. The 3' ends of the restriction fragments, however,
are not extended by incubation with DNA polymerase. Instead, the
strands ligated at their 5' ends to adaptors are melted apart
from their unextended complements and hybridized to a binary
array. The array contains immobilized oligos that are pre-
hybridized with shorter complementary 5'-phosphorylated oligos
that cover (mask) the immobilized oligos except for a segment
which includes a variable region and a region complementary to
the portion of the restriction site remaining at the fragments'
(unrestored) 3' end. The masked region includes the rest of the
restriction site and any other constant sequence, such as may be
included in a priming region. Hybridization is carried out under
conditions that promote the formation of only perfectly matched
hybrids which are the length of the unmasked segment of the
immobilized oligo. After washing away the unbound strands, the
strands that remain bound are ligated to the masking oligos by
incubation with DNA ligase. The correctly bound strands thus
acquire a priming region at their 3' end, in addition to the
priming region they already have at their 5' end. The two
priming regions preferably correspond to different primers. The
array is then washed under appropriately stringent conditions to

-35-

remove all nucleic acids except the immobilized oligos and the
ligated strands hybridized to them.

    1.3.   Sorting RNAs according to their terminal sequences --
    Mature eukaryotic mRNAs share structural features that can
help in their manipulation using arrays. All have a "cap"
structure on their 5' end, and most also possess a 3'-terminal
poly(A) tail, which is attached posttranscriptionally by a
poly(A) polymerase. Because there are usually no long oligo(A)
tracts in the internal regions of cellular RNAs, the poly(A) tail
can serve as a naturally occurring terminal priming sequence in
sorting. The size of mRNAs (several thousand nucleotides in
length) allows them to be amplified and analyzed directly,
without prior cleavage into fragments.
    There are known methods for preparing essentially undegraded
total cellular RNA. Total cellular RNA is converted into com-
plementary DNA (cDNA) using an oligo(dT) primer and a reverse
transcriptase or *Thermus thermophilus* DNA polymerase. Then,
omitting second strand synthesis, single-stranded cDNAs (which
possess oligo(dT) extensions at their 5' end and variable 3'
termini) are sorted according to their 3'-termini on a sectioned
binary array and are ligated there to pre-hybridized adaptors of
a predetermined sequence that are complementary to the immobi-
lized oligos' constant sequence, and that introduce into a cDNA
molecule the 3'-terminal priming site. The cDNA is amplified,
using two primers for PCR: oligo(dT) and an oligo complementary
to the adaptor.

    2.   Preparing partial strands of nucleic acids on oligo-
nucleotide arrays --
    There are two aspects to this procedure: first, the genera-
tion of partial strands (partials), and second, the sorting of
partials according to their terminal oligo segments. All of the
embodiments described below are based on the following principle:
in generating partials from a strand, one of the original strand
ends is preserved (it will be referred to as the "fixed" end),
whereas the other end is truncated to a different extent in the

-36-

various partials (it will be referred to as the "variable" end).
Although either the 5' or the 3' end of the original strand can
serve as the fixed end, it is preferable that the 5' end be
fixed. If amplification of sorted partials is desirable, it is
preferable that the 5' end of the original strand, i.e., the
fixed end, be provided with a priming region prior to partialing
by any of the methods described above, and that partialing be
carried out on a sectioned array. Either an individual strand or
a mixture of strands can be subjected to a partialing; however,
if the mixture is very complex (such as a restriction digest of a
large genome), it is desirable that the mixture first be sorted
into less complex groups of strands, as described above. The
groups of strands used for preparing partials should essentially
be devoid of contaminating strands; therefore, sorting by ter-
minal sequences is preferable for the preliminary sorting. If
preliminary sorting is performed, the strands will already
contain terminal priming regions necessary for amplification of
the partials. Partialing can be performed on either DNA or RNA,
the final product being either DNA or RNA, in either a double-
stranded or a single-stranded state.

    2.1. Methods employing enzymatic cleavage of DNA frag-
ments --
    The purpose of the cleavage is to produce a set of partials
of every possible length; therefore, DNA should be cleaved as
randomly as possible, and to the extent that there is approxi-
mately one cut per strand. Deoxyribonuclease I (DNase I) cleaves
both double-stranded and single-stranded DNA; however, double-
stranded DNA is preferable as the starting material for preparing
partials because of its essentially homogeneous secondary struc-
ture, so that every segment of a DNA molecule is equally acces-
sible to cleavage. Double-stranded DNA fragments are produced as
a result of "symmetric" PCR that can be carried out when sorting
strands. An advantage of using DNase I is that it produces
fragments with 5'-phosphoryl and 3'-hydroxyl termini, that are
suitable for enzymatic ligation.

-37-

After cleavage of the double-stranded DNA fragments, DNase
is removed, e.g., by phenol extraction. The (partial) strands
are then melted apart and are hybridized to a sectioned binary
array, wherein the immobilized oligos are pre-hybridized with
shorter complementary 5'-phosphorylated oligos of a constant
sequence that cover (mask) the immobilized oligos except for a
segment that consists of a variable sequence. Hybridization is
carried out under conditions that favor the formation of per-
fectly matched hybrids of a length that is equal to the length of
the unmasked (variable) segment of the immobilized oligo, and
that minimize the formation of imperfectly matched hybrids.
After washing away unbound strands, the bound strands are ligated
to the masking oligos by incubation with a DNA ligase. The
ligated masking oligos will themselves serve as the second
(3'-terminal) priming region of a partial strand. (All the
partials of a strand will share the same 5' priming sequence that
had been introduced into the strand before generation of the
partials). If restriction fragments are to be partialed that
possess some restriction site at their termini and do not possess
this site internally, it is preferable that the 3' terminal
priming region added to the partials include that site. This
increases the specificity of terminal priming during subsequent
amplification of the partials by PCR. Subsequent extension,
washing, and amplification steps are as described in Example 1.1.
If the partials are prepared for the purpose of sequence deter-
mination, asymmetric PCR can be performed. Alternatively, an RNA
polymerase promoter sequence can be included in one of the two
primers, and amplified DNA is then transcribed to produce multi-
ple single-stranded RNA copies of one of the two complementary
partial strands.

2.2. Methods employing chemical degradation of DNA --
These methods are applicable to both double-stranded and
single-stranded nucleic acids. Chemical degradation is, in most
cases, essentially random. It can be performed under conditions
that destroy secondary structure, and the small size of the

-38-

modifying chemicals makes the chemicals readily accessible to
nucleotides in secondary structures.

Both base-nonspecific reagents and base-specific reagents
can be used. In the latter case, after base-specific cleavage is
performed separately with several portions of the sample, the
portions are mixed together to form a set of all possible partial
DNA lengths. The main drawback to chemical cleavage is that the
location of the terminal phosphate groups on the fragments is
opposite to what is required for enzymatic ligation: 5'-hydroxyl
and 3'-phosphoryl groups are produced in most cases. To overcome
this problem, enzymatic dephosphorylation of 3' ends can be
carried out.

2.3. Method of preparing partials directly on a sectioned
array, without prior degradation of nucleic acids --
In this embodiment, the generation of partials and their
sorting according to the identity of the sequences at their
variable ends occur essentially in one step. First, a strand or
a group of strands (if double-stranded nucleic acid is used as a
starting material, the complementary strands are first melted
apart), is directly hybridized to a sectioned ordinary array,
whose oligos only comprise variable sequences of a pre-selected
length, and that are immobilized by their 5' termini. Optimally,
hybridization is carried out under conditions in which hybrids
can only form whose length is equal to the length of the immobi-
lized oligo. If the array is comprehensive, then a hybrid is
formed somewhere within the array for every oligo that occurs in
a DNA's sequence. After hybridization, the entire array is
washed and incubated with an appropriate DNA polymerase in order
to extend the immobilized oligo, using the hybridized strand as a
template. Each product strand is a partial (complementary) copy
of the hybridized strand. Each partial begins at the place in
the strand's sequence where it has been bound to the immobilized
oligo and ends at the priming region at the 5' terminus of the
strand. If a priming region has not been introduced at the
strand's 5' end before partialing, it can be generated at this
step, after the hybrids that have not been extended, are elimi-

-39-

nated by washing. This can be done either by ligating the 5' end
of the bound strand to a single-stranded oligoribonucleotide
adaptor, or by tailing the immobilized partial copy with a
homopolynucleotide. The entire array is vigorously washed under
conditions that remove the original full-length strands and
essentially all other material not covalently bound. Subsequent
amplification of the immobilized partials can be carried out in
different ways, dependent on whether it is desired to use linear
or exponential amplification.

Exponential copying results in the generation of partials
and their complements. For a strand to be exponentially ampli-
fied by PCR, both of its termini should be provided with a
priming region, preferably different priming regions. The
immobilized (complementary) partial contains only one (3'-
terminal) priming region, and a complementary copy produced by
linear copying would also have only one priming region (on its 5'
end). For RNA copies to have a priming region at their 5' ends,
the immobilized partial should have been provided with an RNA
polymerase promoter downstream of its 3' terminal priming region
using the methods described herein. The second priming region
that is needed for exponential amplification can be introduced at
the 3' ends of the complementary copies as follows.

(a) The 3' termini of RNA copies can then be ligated to
oligoribonucleotide or oligodeoxyribonucleotide adaptors which
are phosphorylated at their 5' end and whose 3' end is blocked.
Exponential PCR can be performed by utilizing the two primers
that correspond to the two priming regions, and then incubating
with Tth DNA polymerase.

(b) If the amplified copies are DNA, they can be trans-
ferred, such as by blotting, (after melting them free of the
immobilized partial) onto a binary array that is a mirror copy of
the first array in the arrangement of the variable segments of
its immobilized oligos. The constant segments of this binary
array are pre-hybridized to masking oligos whose ligation to the
3' termini of the transferred DNAs (by DNA ligase) results in
generation of the second priming region to permit exponential
PCR.

-40-

In methods (a) and (b), both priming regions preferably contain, when applicable, the recognition sequence of the restriction endonuclease that was used to digest the genomic DNA before full-length strand sorting, and which had thus been substantially eliminated from the strands' internal regions.

(c) If partials are surveyed only for oligos that occur in one complementary strand (such as detecting only parental oligos), either only one of the two different primers should be labeled, or the primers should be labeled differently. It is also possible to use labeled substrates during asymmetric PCR.

3. Surveying oligonucleotides with binary arrays --

Surveying oligo content can be carried out in the different embodiments of the invention by hybridization of strands (or partials) to an ordinary array, followed by detection of those hybridized. However, the signal-to-noise ratio is not high enough to always avoid ambiguous results. The most significant problem is inability to sufficiently discriminate against mis-matched basepairs that occur at the ends of hybrids. That hampers analysis of complex sequences. The use of binary arrays helps to overcome this problem.

Binary arrays are also useful for surveying longer oligos than are easily surveyed on an ordinary array (e.g., signature oligos) without increasing the size over that of an ordinary array.

Immobilized oligos in a binary survey array can have either free 5' or 3' ends, and the constant segment can be either upstream or downstream. In most cases, it is preferable that the 3' ends of immobilized oligos be free, and that their constant segments be upstream.

Surveying can utilize sectioned arrays. However, the use of plain arrays is preferable because they are less expensive and more amenable to miniaturization. The following methods are based on the use of plain binary arrays and involve fragmentation of the strands or partials prior to surveying.

-41-

3.1. Comprehensive surveys of DNA strands --

Every oligo present in a strand or in a partial, or in a
group of strands or partials, is surveyed. If a survey of
partials is performed in order to establish nucleotide sequences,
it is preferable that each partial be represented by the same
sense copies. Thus, there should be only one of the complemen-
tary strands in a sample or the complementary strands should be
differentiable, e.g., one strand should produce either no de-
tectable signal or a weaker signal. This can be accomplished by
amplifying the partials linearly or by the use of asymmetric PCR.

DNA strands (or partials) to be surveyed are preferably
digested with nuclease S1 under conditions that destabilize DNA
secondary structure. The digestion conditions are chosen so that
the DNA pieces produced are as short as possible, but at the same
time, most are at least one nucleotide longer than the variable
segment of the oligos immobilized on the binary array. If the
surveyed strands or partials have been previously sorted and
amplified on a sectioned array, this degradation procedure can be
performed simultaneously in each well of that array. Alterna-
tively, if it is desired to store that array as a master for
later use, the array can be replicated by blotting onto another
sectioned array. The DNA is then amplified within the replica
array by (asymmetric) PCR prior to digestion with nuclease S1.

After digestion, the nuclease is inactivated by, for ex-
ample, heating to 100°C, and the DNA pieces are hybridized to an
array whose immobilized oligos' constant segments are pre-
hybridized to 5'-phosphorylated complementary masking oligos.
Preferably, the constant segment contains a restriction site that
has been eliminated from the internal regions of the strands
prior to sorting and is long enough so that its hybrid with the
masking oligo is preserved during subsequent procedures.

The array is incubated with DNA ligase to ligate the masking
oligos to only those hybridized DNA strands (or partials) whose
3' terminal nucleotide is immediately adjacent to the 5' end of
the masking oligo, and matches its counterpart in the immobilized
oligo. DNA ligase is especially sensitive to mismatches at the
junction site.

-42-

After all non-ligated DNA pieces have been washed away under much more stringent conditions that were used during hybridiza-tion, the immobilized oligos are extended by incubation with a DNA polymerase, preferably by only one nucleotide, using the protruding part of the ligated DNA piece as a template, and preferably using the chain-terminating 2',3'-dideoxynucleotides as substrates. Extension is only possible, if the 3'-terminal base of the immobilized oligo forms a perfect basepair with its counterpart in the hybridized DNA piece. The use of the dideoxy-nucleotides ensures that all hybrids are extended by exactly one nucleotide and that all are of the same length. The array is then washed under conditions sufficiently stringent to remove unextended hybrids.

        3.2.   Detection of hybrids --
        Hybrids can be detected by a number of different means. Unlabeled hybrids can be detected by using surface plasmon resonance techniques, which currently can detect 10^8 to 10^9 hybrid molecules per square millimeter. Alternatively, hybrids can be conventionally labeled, such as with radioactive or fluorescent groups. Fluorescent labels are convenient.
        To ensure the lowest level of background labeling, it is preferable to label hybrids in a manner such that its detection is dependent on the success of both a ligation and an extension step. This can be accomplished within the scheme of oligo surveying by labeling the masking oligos, and the 2',3'-dideoxy-nucleotides used for the extension with fluorescent dyes possess-ing different emission spectra. The array can then be scanned at different wavelengths, corresponding to the emission maxima of the two dyes, and only signals from those areas that emit fluo-rescence of both colors are taken as a positive result.
        After hybrids are extended (concomitant with labeling) and edited, the array is thoroughly washed to remove unincorporated label, destroy unextended hybrids, and discriminate one more time against mismatched hybrids that might have remained. A preferred method is to wash the array at steadily increasing temperature, with the signal from each area being read at a pre-determined

-43-

time, when the conditions ensure the highest selectivity for the
particular hybrid that forms in that area. Other conditions
(such as denaturant and/or salt concentration) can also be
controlled over time. The fluorescence pattern can be recorded
at predetermined time intervals with a scanning microfluorometer,
such as an epifluorescence microscope.

        4.   Determination of the nucleotide sequences of strands in
a mixture when each strand possesses at least one oligo that does
not occur in any other strand in the mixture --
        Figures 8 to 11 depict the determination of the sequences of
two mixed strands using the methods of the invention. The
example demonstrates the power of the invention to identify all
the oligos present in a strand (i.e., its strand set) when it
possesses at least one oligo that does not occur in any other
strand in the mixture. In particular, the example demonstrates:
(a) how the data obtained by surveying the partial strands
generated from a mixture of strands and sorted by their variable
termini (i.e., the upstream subset of each address) and the
inferred downstream subset of each address (which together form
the indexed address sets) are used to construct the unindexed
address sets; and (b) how the unindexed address sets are compared
to each other to identify prime sets. The example also demon-
strates how the oligos contained in a strand set are assembled
into the sequence of the strand, even though the primary data is
obtained from a mixture. In particular, the example demon-
strates: (a) how oligos in a strand set are assembled into
sequence blocks; (b) how the contents of the indexed address sets
are filtered so that only information pertaining to the oligos in
a particular strand set remains; (c) how this filtered data is
re-expressed in terms of the sequence blocks that are contained
in that particular strand; (d) how information in the resulting
"block sets" is used to identify those blocks that definitely
occur only once in the strand ("unique blocks") and to identify
those that can potentially occur more than once; (e) how informa-
tion in block sets of unique blocks is used to determine the
relative order of the blocks that occur only once in the strand;

-44-

(f) how the information in the block sets limits the positions at
which the other blocks can occur (relative to other blocks); and
(g) how a consideration of the sequences at the ends of blocks,
in combination with a consideration of the relative positions of
the blocks, leads to the unambiguous determination of the com-
plete sequence of the strand. This example also illustrates:
(a) how oligos that occur more than once in a strand are identi-
fied and located within the sequence, even though the survey data
contain no information as to the number of times a particular
oligo occurs in a partial or a mixture of partials having the
same terminal oligo; and (b) how the sequences of different
strands in a mixture can be determined separately, despite the
fact that many of the oligos occur in more than one strand.

     Figure 8a shows the sequences of two short strands (parental
strands) that are assumed to be present in a mixture (with no
other strands). It is assumed that complete sets of partials
have been generated from this mixture, and that each set of
partials has been separately surveyed, with the partials sharing
the same address oligo being surveyed together. For the purpose
of illustrating the method of analyzing the data, it is assumed
that the address oligos and the surveyed oligos are three nucleo-
tides in length. In practice, longer oligos should be used.
However, for illustration it is easier to comprehend an example
based on trinucleotides. The same methods of analyzing the data
apply when longer oligos are surveyed, when much longer strands
are in the mixture, and when the mixture contains many more
strands.

     Figure 8b shows the upstream subsets determined by surveying
and the downstream subsets inferred (i.e., Figure 8b shows
indexed address sets). The address oligos (bold letters) are
listed vertically in the center of the diagram. The oligos
listed horizontally to the left of each address oligo are those
oligos that were detected in a survey of the partials at that
address (the upstream subset). The oligos listed horizontally to
the right of each address oligo are those inferred from the
upstream subsets to occur downstream of that address oligo (the
downstream subset). For example, oligo "ACC" is contained in the

-45-

upstream subset of the address oligo "CCT". This means that
oligo "CCT" occurs downstream of oligo "ACC" in at least one
strand in the mixture. Therefore "CCT" is inferred to be in the
downstream subset of address set "ACC". The remaining downstream
oligos in all of the address sets are similarly inferred. Note
that an address oligo is a member of its own upstream and down-
stream subsets.

After the indexed address sets of all addresses in the
parental strands have been determined (as shown in Figure 8b),
the information is organized into unindexed address sets (Figure
8c), having no division into downstream and upstream subsets, but
merely listing, for each address oligo, those oligos that occur
in either the upstream or downstream subset (or in both). In
Figure 8c, the address oligos (bold letters) are listed verti-
cally on the left side of the diagram. Note that the address
oligo is a member of its own unindexed address set.

Unindexed address sets are grouped together according to the
identity of the oligos they contain (Figure 8d). Unindexed
address sets that contain an identical set of oligos are grouped
together. It can be seen that three groups of address sets are
formed in this example. The groups are identified by the Roman
numerals (I, II, and III). The address oligos of each group (for
example, CTA, GTC, and TCC in group II) always occur together in
a strand and can occur together in more than one strand.

Each group of identical address sets is then compared to all
other groups of identical address sets to see if its common
address set appears to be a prime by seeing whether any other
address set is a subset of it. For example, in Figure 8d, the
address set common to group III is not a prime address set,
because the address set common to group I is a subset of the
address set common to group III. However, the address set common
to group I and the address set common to group II appear to be
prime address sets.

Each putative prime address set is then tested to see if it
is a strand set by examining all the address sets that contain
all of the oligos that are present in it. For example, in Figure
9a, all the address sets that contain all the oligos present in

-46-

the putative prime address set common to group I are listed
together (namely the address sets contained in groups I and III).
The address oligos are shown in bold letters on the left side of
the diagram, and the groups are identified by Roman numerals.
The address set common to group I is indeed a prime address set
(and therefore it contains a single strand set) because a list of
the eleven oligos that are found in every address set in the
diagram (they are seen as full columns) is identical to the list
of eleven addresses on the left side of the diagram. Similarly,
Figure 8b shows why the address set common to group II is also a
prime set. The twelve oligos common to every address set in the
diagram are all found in the list of twelve addresses on the left
side of the diagram. Had either of these putative prime address
sets not turned out to be a prime set (by the criterion described
above), then it would have been identified as a pseudo-prime
address set, and further analysis would have been required to
decompose it into its constituent strand sets.

Once the strand sets in a mixture have been identified, the
oligos in each strand set can be assembled into the strand
sequence in a series of steps, as illustrated in Figure 10 (which
utilizes the strand set determined in Figure 9a).

First the oligos in the strand set are assembled into
sequence blocks. A sequence block contains one or more uniquely
overlapping oligos. Two oligos of length n, uniquely overlap
each other if they share an identical sub-sequence that is $n-1$
nucleotides long and no other oligos in the same strand set share
that sub-sequence. For example, for the strand set shown in
Figure 10a, the oligos "CAT" and "ATG" share the sub-sequence
"AT" which does not occur in other oligos. These two oligos
therefore uniquely overlap to form the sequence block "CATG", as
shown in Figure 10b. Similarly, oligo "TGG" uniquely overlaps
oligo "GGT" by the common sub-sequence "GG", and oligo "GGT" also
uniquely overlaps (on its other end) oligo "GTA" by the common
sub-sequence "GT". Thus, the three oligos ("TGG", "GGT", and
"GTA") can be maximally overlapped to form sequence block
"TGGTA". In forming sequence blocks, the following rule is
adhered to: two oligos can be included in the same block if they

-47-

are the only oligos in the strand set to possess their common
sub-sequence. Thus, "ATG" does not uniquely overlap "TGG",
because the strand set contains a third oligo, "TTG", that shares
the common sub-sequence "TG". If, following these rules, an
oligo does not uniquely overlap any other oligo, then a sequence
block consists of only that oligo. For example, "TAA" forms its
own block. Following the above rules, the eleven oligos that
occur in strand set A can be assembled into four sequence blocks.

Second, the data contained in the indexed address sets shown
in Figure 8b are filtered to remove extraneous information that
does not pertain to strand set A. Figure 10c shows the resulting
filtered address sets. All address sets whose address oligo is
not one of the oligos in strand set A are eliminated. In addi-
tion, all oligos that are not members of strand set A are removed
from the upstream and downstream subsets of the remaining address
sets. The resulting filtered address sets are then grouped
together according to the oligos that are contained in each
block. For example, the filtered address sets for address oligos
"CAT" and "ATG" have been grouped together in Figure 10c because
these two oligos are contained in sequence block "CATG". In
Figure 10c, the address oligos found in the same block are
identified by rectangular boxes. In addition, oligos that occur
in the same block are grouped together within each upstream and
downstream subset.

Third, the filtered address sets are converted into block
sets, as shown in Figure 10d. In a block set, the information
from different address sets is combined. Instead of a different
horizontal line for each filtered address set that pertains to a
particular block, the information in all of the address sets that
pertain to that particular block is combined into a single
horizontal line. For example, in Figure 9c, five different
filtered address sets pertain to sequence block "TACCTTG". In
Figure 10d, these five lines are combined into a single line in
which the address oligos are replaced by an "address block",
shown as "TACCTTG" surrounded by a bold box. Similarly, the
upstream oligos are replaced by upstream blocks, and the down-
stream oligos are replaced by downstream blocks. In substituting